На правах рукописи

Крымова Екатерина Александровна

# Сплайны в задачах интерполяции и регрессионного анализа гауссовских процессов и гладких функций.

05.13.17 – Теоретические основы информатики.

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата физико-математических наук

Москва – 2013

Работа выполнена в секторе 5 Интеллектуального анализа данных и моделирования Федерального государственного бюджетного учреждения науки Института проблем передачи информации им. А.А. Харкевича Российской академии наук.

Научный руководитель: *доктор физико-математических наук,*
***Голубев Георгий Ксенофонтович***

Официальные оппоненты: *доктор физико-математических наук,*
***Бурнашев Марат Валиевич,***
*лаборатория №1 Института проблем передачи информации им. А.А. Харкевича РАН, ведущий научный сотрудник*

*доктор физико-математических наук, профессор,*
***Назин Александр Викторович,***
*Институт проблем управления им. В.А. Трапезникова РАН, ведущий научный сотрудник*

Ведущая организация: *федеральное государственное бюджетное учреждение науки Вычислительный центр им. А.А. Дородницына Российской академии наук.*

Защита состоится «____» _____ 2013 г. в _____ часов на заседании диссертационного совета Д 212.156.04 при Московском физико-техническом институте (государственном университете) по адресу: 141700, г. Долгопрудный, Московская обл., Институтский пер., д. 9, ауд. 204 нового корпуса.

С диссертацией можно ознакомиться в библиотеке Московского физико-технического института (государственного университета).

Автореферат разослан «____» _____ 2013 г.

Ученый секретарь
диссертационного совета Д 212.156.04
к.ф.-м.н. Стрыгин Л. В.

## Общая характеристика работы

**Актуальность работы.** Диссертация посвящена задачам математической теории интерполяции стационарных гауссовских процессов и оценивания гладких функций регрессии. При этом существенное внимание уделяется интерполяционным методам и методам оценивания, основанным на сплайнах.

Задачи интерполяции функций естественным образом возникают в различных областях прикладной математики, а методы интерполяции широко используются в многочисленных инженерных приложениях. Как правило, задача интерполяции заключается в восстановлении неизвестной функции по ее значениям, заданным на дискретном множестве точек. Очевидно, что в общем случае точно восстановить функцию во всех точках невозможно и поэтому основной целью является поиск методов интерполяции, обеспечивающих минимально возможную ошибку интерполяции. Понятно также, что как ошибка интерполяции, так и метод интерполяции критически зависят от имеющейся априорной информации об интерполируемой функции. Во многих случаях довольно естественно предполагать, что интерполируемая функция является гладкой. При этом естественно возникает неформальная задача о том, как оптимальным образом трансформировать интуитивное понятие гладкости в метод интерполяции.

Классические подходы к интепроляции гладких функций связаны с интерполяциями с помощью полиномов. Они разрабатывались Лагранжем, Ньютоном, Стирлиногом и др. Хорошо известно, что интерполяция полиномами становится крайне неустойчива при возрастании числа наблюдений (феномен Рунге), к тому же нет возможности контролировать степень гладкости получающейся интерполяции. Именно поэтому возникла идея использования интерполяционных локальных полиномов невысокой степени. Для снижения погрешности интерполяции отрезок наблюдения функции разбивается на несколько отрезков и на каждом из них строится интерполяционный локальный полином, затем полиномы гладко сшиваются. Степень локального полинома чаще выбирается из априорного представления о

гладкости функции. Эта идея по-разному реализована в методах кусочно-гладкой интерполяции Лагранжа, Эрмита (при заданных производных в точках наблюдения), сплайнах. Отметим также, что локальные полиномы используются в барицентрическом методе дробно-рациональной интерполяции Флоатера. Однако точность этого метода при нерегулярном расположении точек чаще всего неудовлетворительна.

Принципиально иной подход к задаче интерполяции основан на использовании вероятностных моделей для интерполируемой функции. Наиболее часто используемый на практике, в особенности, в геостатистике, метод кригинга использует предположение о том, что наблюдаемая функция является реализацией гауссовского процесса с ковариационной функций из некоторого заданного параметрического семейства. К сожалению, практически никогда нет уверенности в том, что интерполируемый процесс принадлежит выбранному классу. Также при построении интерполяции приходится оценивать параметры неизвестной ковариационной функции, что приводит к невыпуклой задаче оптимизации.

Среди многочисленных методов интерполяции функций, используемых на практике, сплайны занимают особое место. Это прежде всего обусловлено тем, что они

- позволяют хорошо интерполировать гладкие функции;
- имеют простую и ясную физическую интерпретацию. В частности, кубический сплайн описывается формой тонкой гибкой линейки, проходящей через заданные точки;
- допускают исключительно быстрые алгоритмы для их вычисления.

Эти свойства сплайнов были замечены и использованы инженерами очень давно, по-видимому, первое упоминание о сплайнах содержится в книге XVIII века А.-Л. Дюамеля дю Монсо.

Широкое использование сплайнов на практике требует их всестороннего теоретического обоснования. Поэтому, в частности, возникает задача сравнения интерполяции сплайнами с минимаксными интерполяциями гладких стационарных гауссовских процессов и функций из соболевских классов. Кроме того, в инженерных приложениях часто требуется не только построить хороший метод интерполяции, но и оценить точность интерполяции, которую он может обеспечить. К сожалению, в рамках классической теории функциональной интерполяции последняя задача не имеет решения. Ее решение становится возможным при некоторой дополнительной априорной информации об интерполируемой функции. В качестве такой информации может служить гипотеза о том, что функция представляет собой реализацию гауссовского процесса.

Метод решения задачи о вычислении минимаксной интерполяции и ее точности для соболевского класса гладких функций идейно близок к методу, предложенному М. С. Пинскером для асимптотически точного вычисления минимаксного риска фильтрации квадратично-интегрируемых сигналов. Точное аналитическое решение задачи минимаксной интерполяции возможно, если значения функции задаются на бесконечной равномерной решетке. В этом случае можно осуществить переход в спектральную область с помощью преобразования Фурье (см., например, статьи Г. К. Голубева , Г. К. Голубева и М. Нусбаума). При этом нижние границы для ошибки интерполяции получаются на основе решения хорошо известной задачи об интерполяции стационарных стохастических последовательностей, которая была детально изучена в работах А. Н. Колмогорова и Н. Винера.

Задача интерполяции является предельным случаем задачи оценивания функции регрессии при уровне шума, стремящемся к нулю. Поэтому в диссертации наряду с задачами интерполяции рассматриваются задачи восстановления функции регрессии с помощью сглаживающих сплайнов. При использовании сглаживающих сплайнов в случае ненулевого уровня шума возникает проблема выбора параметра сглаживания. Часто он находится с помощью метода GCV, который является одним из вариантов метода несмещенного оценивания риска. Для рис-

ка оценок, полученных с помощью метода несмещенного оценивания риска, А. Кнайп получил очень хорошие верхние границы, равномерные по всем оцениваемым функциям регрессии, которые часто называются в современной математической статистике оракульными неравенствами.

Очевидно, что выбор наилучшей оценки из заданного семейства оценок является частным случаем поиска наилучшей выпуклой комбинации оценок из этого семейства. Такой метод построения оценок называется агрегацией. Первые подходы к агрегации оценок были основаны на разбиении наблюдений на две независимые части. При этом оценки строились по одной части наблюдений, а наилучшая выпуклая комбинация вычислялась по другой. Этот подход был разработан независимо А. Немировским и О. Катони. Разбиение выборки на две части неизбежно влечет потери статистической информации, содержащейся в наблюдениях, и на практике его естественно стараются избежать. С математической точки зрения деление выборки приводит к тому, что получающиеся верхние границы для риска агрегированной оценки оказываются хуже границ Кнайпа. Существенный прогресс в методах агрегации, не использующих разбиение выборки, был достигнут в работе Г. Леюнга и А. Баррона для метода экспоненциального взвешивания. Дальнейшее развитие методов этой работы сделано Г.К. Голубевым для агрегации проекционных оценок. Отметим также, что несколько иные результаты для метода агрегации функций из словаря с помощью метода экспоненциального взвешивания в задаче восстановления функции регрессии получены недавно Ф. Риголле и А. Цыбаковым, А. Далаляном и Ж. Салмоном.

Поэтому **цели** данной работы состоят в том, чтобы:

- математически обосновать близость интерполяционных сплайнов к наилучшим методам интерполяции функций из соболевских классов;
- разработать метод контроля точности для интерполяционных сплайнов;
- получить оракульные неравенства для экспоненциальной агрегации сглаживающих сплайнов, которые улучшают неравенство Кнайпа.

В соответствии с перечисленными целями были определены **задачи** исследования:

1. Вычислить ошибку минимаксной интерполяции гауссовских процессов из соболевских классов и сравнить его с ошибкой интерполяции сплайнами.

2. Рассмотреть задачу минимаксной интерполяции гладких функций на равномерной решетке со случайным сдвигом.

3. Предложить и обосновать метод контроля точности сплайновой интерполяции на основе эквивалентности сплайнов и оптимальной интерполяции для гауссовских стационарных процессов со специальными спектральными плотностями.

4. Доказать новые оракульные неравенства для задачи оценивания функции регрессии с помощью метода экспоненциального взвешивания, улучшающие известные результаты.

5. Экспериментально сравнить метод экспоненциального взвешивания с методом несмещенного оценивания риска.

**Общая методика исследования.** Для решения поставленных задач в работе используются методы математической статистики, теории случайных процессов, теории вероятности, аппарат анализа Фурье.

**Научная новизна** результатов, полученных в диссертации, заключается в том, что предложен новый метод оценивания качества интерполяции методом сплайнов. Основываясь на вероятностных свойствах несмещенной оценки риска, доказаны новые оракульные неравенства для метода экспоненциального взвешивания упорядоченных оценок. Причем остаточный член в полученных оракульных неравенствах улучшен по сравнению с результатом Кнайпа.

**Практическая значимость.** Практическая значимость диссертационной работы определяется широким использованием предложенного метода контроля

качества сплайнов, реализованного в программном продукте Macros компании Datadvance, в частности, для решения ряда прикладных задач концерна EADS.

**На защиту выносятся следующие результаты:**

1. Показано, что риск интерполяции сплайнов близок к риску минимаксной интерполяции гладких стационарных гауссовских процессов.

2. Вычислен риск минимаксной интерполяции гладких функций на равномерной решетке со случайным сдвигом. Полученный риск равен минимаксному риску интерполяции гладких гауссовских стационарных процессов.

3. Предложен метод контроля точности интерполяции сплайнами. Показано, что для определенного класса процессов предложенный метод является хорошей оценкой для реальной ошибки интерполяции.

4. Задачи восстановления функции регрессии с помощью сглаживающих сплайнов сведены к задаче оценки зашумленного вектора при заданном множестве упорядоченных оценок. Для метода экспоненциального взвешивания упорядоченных оценок выведены новые оракульные неравенства.

5. Проведены численные эксперименты, которые показали, что в случае, когда отношение риска оракула к дисперсии шума мало, экспоненциальное взвешивание позволяет получить оценку с меньшим риском.

**Апробация работы.** Результаты работы докладывались и обсуждались на следующих конференциях:

- 18-я European Young Statisticians Meetings (2013, Осиек, Хорватия);
- 43-rd Probability Summer school (2013, Сент-Флур, Франция);
- 9-я Международная конференция «Интеллектуализация обработки информации» (2012, Будва, Черногория);

- Международная конференция по вероятности и предсказательному моделированию (2012, Москва, Россия);
- Международная конференция молодых ученых «Информационные Технологии и Системы» (2012, Петрозаводск, Россия; 2013, Калининград, Россия);
- 55-я Всероссийская научная конференция Московского физико-технического института (2012, Долгопрудный, Россия).

Также результаты работы обсуждались на семинарах Лаборатории структурных методов анализа данных в предсказательном моделировании МФТИ (2012, 2013).

**Публикации.** Основные результаты по теме диссертации изложены в восьми печатных изданиях, из которых [1–3] изданы в журналах, рекомендованных ВАК.

**Личный вклад автора.** Содержание диссертации и основные положения, выносимые на защиту, отражают персональный вклад автора в опубликованные работы. Подготовка к публикации полученных результатов проводилась совместно с соавторами, причем вклад диссертанта был определяющим.

**Структура и объем диссертации.** Диссертация состоит из введения, 3 глав, заключения и библиографии. Общий объем диссертации 97 страниц, включая 15 рисунков. Библиография включает 65 наименований.

**Благодарности.** Автор благодарен своему научному руководителю Георгию Ксенофонтовичу Голубеву за постановки задач, плодотворные обсуждения, за постоянную поддержку и участие.

Работа выполнена при поддержке Лаборатории структурных методов анализа данных в предсказательном моделировании, МФТИ, грант правительства РФ дог. 11.G34.31.0073.

## Содержание работы

**Во Введении** обоснована актуальность диссертационной работы, сформулирована цель и аргументирована научная новизна исследований, показана практическая значимость полученных результатов, представлены выносимые на защиту положения.

**В первой главе** рассматривается задача интерполяции неизвестной функции $f(x),\ x \in \mathbb{R}$. А именно, с помощью данных $(X, Y) = \{X_k,\ Y_k = f(X_k),\ k = 1, \ldots, n\}$ необходимо восстановить значение функции $f(x)$ в некоторой заданной точке $x \in (0, 1)$. Здесь и далее для простоты предполагается, что все точки $X_k$ различны, упорядочены $X_1 < X_2 < \ldots < X_n$ и принадлежат отрезку $[0, 1]$.

**В первом разделе** приводится краткий обзор наиболее часто используемых на практике методов интерполяции.

**Во втором разделе** изучается интерполяция стационарного гауссовского процесса со спектральной плотностью $F_\alpha(\omega) = Q/(\omega^{2m} + \alpha^{2m})$, где $Q$, $\alpha$ – положительные, как правило, неизвестные параметры, $m \geq 1$ – известное целое число.

Хорошо известно, что наилучшая в смысле квадратичного критерия качества интерполяция гауссовского процесса со спектральной плотностью $F_\alpha(\omega)$ имеет вид

$$\bar{f}(x, X, Y) = \sum_{k=1}^{n} K_{\alpha,Q,m}(x, X_k) Y_k,$$

где ядро $K_{\alpha,Q,m}(\cdot, \cdot)$ определяется как решение уравнения Винера-Хопфа-Колмогорова

$$\mathbf{E}\left[f(x) - \sum_{k=1}^{n} K_{\alpha,Q,m}(x, X_k) f(X_k)\right] f(X_s) = 0, \tag{1}$$

где $s = 1, \ldots, n$. К сожалению, использовать это уравнение на практике затруднительно, так как параметры $\alpha$ и $Q$, как правило, неизвестны. В принципе, эти параметры можно было бы оценить, например, с помощью метода максимального правдоподобия, но такой подход существенно усложняет вычисление оптимальной интерполяции поскольку приводит к нелинейной задаче оптимизации. Гораздо более робастным и простым подходом является поиск решения уравнения (1) при

$\alpha \to 0$. Заметим однако, что поскольку

$$\lim_{\alpha\to 0} \frac{Q}{\omega^{2m}+\alpha^{2m}} = \frac{Q}{\omega^{2m}} \quad \text{и} \quad \int\limits_{-\infty}^{\infty} \frac{1}{\omega^{2m}}\, d\omega = \infty,$$

случайный процесс со спектральной плотностью $Q\omega^{-2m}$ не существует в обычном смысле. Тем не менее, предельное ядро $K_m(x,X) = \lim_{\alpha\to 0} K_{\alpha,Q,m}(x,X)$ существует. Чтобы найти уравнения, которым оно удовлетворяет, определим функции

$$d_s^{(0)}[x] = |X_s - x|^{2m-1}, \quad d_s^{(j+1)}[x] = \frac{d_{s+1}^{(j)}[x] - d_s^{(j)}[x]}{X_{s+j+1} - X_s}.$$

**Теорема 1.** *Пусть все точки $X_k,\ k = 1,2,\ldots,n$ различны и $n \geq m+1$. Тогда $K_m(x,X_k)$ является решением системы линейных уравнений*

$$\sum_{k=1}^{n} K_m(x,X_k) d_s^{(m)}[X_k] = d_s^{(m)}[x],\ s = 1,\ldots,n-m,$$
$$\sum_{k=1}^{n} K_m(x,X_k) X_k^p = x^p,\ p = 0,\ldots,m-1.$$

Основным результатом раздела является эквивалентность интерполяции сплайнами и оптимальной интерполяции обобщенного гауссовского процесса со спектральной плотностью $\omega^{-2m}$. Интерполяционный сплайн определяется как предел при $\epsilon \to 0$ решения следующей оптимизационной задачи:

$$\bar{S}_{Q,m}^{\epsilon}(x,X,Y) = \arg\min_f \left\{ \frac{1}{2\epsilon^2} \sum_{j=1}^{n} \left[Y_j - f(X_j)\right]^2 + \frac{1}{2Q} \int\limits_0^1 \left[f^{(m)}(x)\right]^2 dx \right\}.$$

Интерполяционный сплайн линеен по $Y_k,\ k = 1,\ldots,n$ и его можно записать в виде

$$\bar{S}_{Q,m}^{\epsilon}(x,X,Y) = \sum_{k=1}^{n} K_{Q,m}^{\epsilon}(x,X_k) Y_k.$$

**Теорема 2.** *Предположим, что все точки $X_j,\ j = 1,\ldots,n$, различны и $n \geq m$. Тогда*

$$\lim_{\epsilon\to 0} K_{Q,m}^{\epsilon}(x,y) = K_m(x,y),$$

*где ядро $K_m(\cdot,\cdot)$ определено в теореме 1.*

**В третьем разделе** рассматривается задача сравнения минимаксного риска интерполяции и интерполяции методом сплайнов в классе гладких стационарных гауссовских процессов. Пусть $f(\cdot)$ — стационарный гауссовский процесс с известной спектральной плотностью $F(\omega)$. Задача состоит в том, чтобы восстановить $f(x)$ на интервале $[0,h]$ по наблюдениям $Y_k = f(X_k)$, где $X_k = kh,\ k = 0, \pm1, \pm2, \ldots$. Поскольку процесс гауссовский и стационарный, то его наилучшая интерполяция имеет вид

$$\bar{f}(x, X, Y) = h \sum_{k=-\infty}^{\infty} K(x - X_k) Y_k,$$

где $K(\cdot)$ симметричное ядро, которое находится из минимизации средне квадратичной ошибки интерполяции. Для интерполяции $\tilde{f}(x, X, Y)$ эта ошибка определяется следующим образом:

$$\sigma^2(\tilde{f}, F) = \frac{1}{h} \int_0^h \mathbf{E}[f(x) - \tilde{f}(x, X, Y)]^2 \, dx.$$

Предположим, что процесс $f(x)$ гладкий, точнее, что он является гуссовским процессом, спектральная плотность которого принадлежит классу $\mathcal{F}^m(L)$, определяемому условием

$$\mathbf{E}[f^{(m)}(x)]^2 \le L.$$

Рассмотрим задачу минимаксной интерполяции гладких процессов со спектральными плотностями из этого класса. С математической точки зрения эта задача состоит в том, чтобы

- вычислить минимаксную ошибку интерполяции

$$R_h^m(L) \stackrel{\text{def}}{=} \inf_{\tilde{f}} \sup_{F \in \mathcal{F}^m(L)} \sigma_h^2(\tilde{f}, F),$$

  где inf вычисляется по всем интерполяциям;

- построить минимаксную интерполяцию $\bar{f}_*(\cdot, X, Y)$, то есть такую, что

$$R_h^m(L) = \sup_{F \in \mathcal{F}^m(L)} \sigma_h^2(\bar{f}_*, F).$$

**Теорема 3.** *Минимаксная ошибка интерполяции вычисляется как*

$$R_h^m(L) = \frac{L}{2}\left(\frac{h}{\pi}\right)^{2m}.$$

*При этом минимаксная интерполяция имеет вид*

$$\bar{f}_*(x, X, Y) = \frac{1}{h}\sum_{k=-\infty}^{\infty} K_*\left(\frac{x - X_k}{h}\right) f(X_k),$$

*где $K_*(\cdot)$ — симметричное ядро, преобразование Фурье которого определяется как*

$$\hat{K}_*(\omega) = \begin{cases} 1, & \omega \in [0, \omega_*), \\ 2^{m-1}(1-\omega)^m, & \omega \in [\omega_*, 1/2), \\ 1 - 2^{m-1}\omega^m, & \omega \in [1/2, 1-\omega_*), \\ 0, & \omega \geq 1 - \omega_*; \end{cases}$$

*здесь $\omega_* = 1 - 2^{-1+1/m}$.*

Чтобы понять, насколько хорошо сплайны интерполируют процессы со спектральными плотностями из класса $\mathcal{F}^m(L)$, вычислим максимальную ошибку интерполяции

$$\sigma_h^2\left[\mathcal{F}^m(L), \bar{S}_m\right] = \sup_{f \in \mathcal{F}^m(L)} \sigma_h^2(f, \bar{S}_m);$$

здесь $\bar{S}_m(x, Y) = \lim_{\epsilon \to 0} \bar{S}_m^\epsilon(x, Y)$ — интерполяционный сплайн, который в силу теорем 1, 2 имеет следующий вид:

$$\bar{S}_m(x, X, Y) = \frac{1}{h}\sum_{k=-\infty}^{\infty} K_m\left(\frac{x - X_k}{h}\right) f(X_k).$$

При этом преобразование Фурье ядра $K_m(\cdot)$ определяется как

$$\hat{K}_m(\omega) = \left[1 + \sum_{k \neq 0}\left(1 + \frac{k}{\omega}\right)^{-2m}\right]^{-1}.$$

Можно показать, что

$$\sigma_h^2\left[\mathcal{F}^m(L), \bar{S}_m\right] = \frac{L}{2}\left(\frac{h}{\pi}\right)^{2m} \max_\omega q_m(\omega),$$

где

$$q_m(\omega) = \frac{1}{2^{2m-1}\omega^{2m}}\left\{\left[1 - \hat{K}_m\left(\frac{\omega}{h}\right)\right]^2 + \sum_{k\neq 0}\hat{K}_m^2\left(\frac{\omega+k}{h}\right)\right\}.$$

Аналитически вычислить максимум функции $q_m(\omega)$ довольно сложно. Можно подсчитать его численно и найти величину $\sigma_h^2\big[\mathcal{F}^m(L), \bar{S}_m\big]/R_h^m(L)$, которая характеризует эффективность интерполяции сплайнами при различных $m$. Оказывается, что ошибка интерполяции сплайнами довольно близка к минимаксной ошибке интерполяции. Например, эффективность для кубических сплайнов ($m = 2$) приблизительно равна 1,35.

**Четвертый раздел** посвящен интерполяции гладких функций, заданных значениями на бесконечной равномерной решетке со случайным сдвигом. Точнее рассмотрим интерполяцию гладких функций из соболевского класса $\mathcal{W}_T^m(L)$, задаваемого условиями

$$\int_{-\infty}^{\infty}[f^{(m)}(x)]^2\,dx \le LT, \quad \operatorname{supp}\{f\} \in [0,T].$$

При этом будем предполагать, что точки $X_k$ расположены на решетке $X_k^\zeta = kh + \zeta,\ k = 0 \pm 1, \dots$ с шагом $h > 0$. Здесь и далее $\zeta$ – случайная величина равномерно распределенная на $[0, h]$.

Задача состоит в том, чтобы восстановить функцию $f(x),\ x \in [0,T]$ при заданных значениях $Y_k^\zeta = f(X_k^\zeta)$. При этом потенциально наилучшее качество интерполяции определяется минимаксной ошибкой интерполяции

$$\rho_h^T(L) = \inf_{\tilde{f}}\sup_{f\in\mathcal{W}_T^m(L)}\mathbf{E}_\zeta\int_0^T[f(x) - \tilde{f}(x, X^\zeta, Y^\zeta)]^2\,dx,$$

где $\mathbf{E}_\zeta$ усреднение по распределению величины $\zeta$, а inf вычисляется по всем интерполяциям.

Следующий результат показывает, что минимаксные интерполяции гауссовских случайных процессов со спектральными плотностями из класса $\mathcal{F}^m(L)$ и функций из соболевского класса $\mathcal{W}_T^m(L)$ очень близки.

**Теорема 1.** *При* $T \to \infty$

$$\frac{\rho_h^T(L)}{T} = (1 + o(1))\frac{L}{2}\left(\frac{h}{\pi}\right)^{2m}.$$

**В пятом разделе** предлагается простой метод контроля точности интерполяции сплайнами. Будем считать, что $X_k = kh,\ k = 0, \pm1, \pm2, \ldots$ и $Y_k = f(X_k)$. Обозначим $\bar{S}_m(x, X, Y)$ интерполяционный сплайн порядка $m$ и соответственно $\bar{S}_{m+1}(x, X, Y)$ интерполяционный сплайн следующего порядка. В качестве оценки для величины реальной ошибки интерполяции

$$\sigma_m^2(x) = [f(x) - \bar{S}_m(x, X, Y)]^2$$

будем использовать величину

$$\bar{\sigma}_m^2(x) = [\bar{S}_{m+1}(x, X, Y) - \bar{S}_m(x, X, Y)]^2.$$

Следующий результат обосновывает этот метод.

**Теорема 4.** *Пусть* $f(x)$ *– стационарный процесс со спектральной плотностью* $F(\omega)$*, имеющей представление* $F(\omega) = \phi(\omega)\omega^{-2m}$*, где* $\phi(\cdot)$ *– положительная четная функция, не возрастающая при* $\omega > 0$*. Тогда существует постоянная* $V_m$ *такая, что*

$$\int_{X_k}^{X_{k+1}} \mathbf{E}\sigma_m^2(x)\,dx \le V_m \int_{X_k}^{X_{k+1}} \mathbf{E}\bar{\sigma}_m^2(x)\,dx. \tag{2}$$

Если интерполируемая функция обладает большой гладкостью, то можно показать, что постоянная $V_m$ в неравенстве (2) близка к 1, точнее $\lim_{m\to\infty} V_m = 1$.

Заметим также, что условие невозрастания функции $\phi(\omega)$ при положительных $\omega$ можно заменить на условие $A\phi_\circ(\omega) \le \phi(\omega) \le B\phi_\circ(\omega)$, где $A, B$ – некоторые положительные постоянные, а $\phi_\circ(\omega)$ симметричная, не возрастающая функция при положительных $\omega$.

**Во второй главе** рассматриваются методы агрегации линейных упорядоченных оценок в задаче восстановления неизвестного вектора по зашумленным

данным. Эти задачи играют принципиально важную роль, в частности, при оценивании функции регрессии с помощью сглаживающих сплайнов.

**В первом разделе** приводится постановка и мотивация задачи оценивания вектора $\mu = (\mu_1, \dots, \mu_n)^\top$ по наблюдениям

$$Y_i = \mu_i + \sigma\xi_i, \quad i = 1, 2, \dots, n, \tag{3}$$

где $\xi_i$ белый гауссовский шум. Для простоты предполагается, что параметр $\sigma > 0$ известен. Основная цель состоит в том, чтобы построить оценку вектора $\mu$ на основе семейства линейных оценок

$$\hat{\mu}_i^h(Y) = h_i Y_i, \ h \in \mathcal{H}, \tag{4}$$

где $\mathcal{H}$ — заданное множество векторов из $\mathbb{R}^n$, которое будет описано ниже.

Риск оценки $\hat{\mu}(Y) = (\hat{\mu}_i(Y), \dots, \hat{\mu}_n(Y))^\top$ измеряется величиной

$$R(\hat{\mu}, \mu) = \mathbf{E}_\mu \|\hat{\mu}(Y) - \mu\|^2,$$

здесь $\mathbf{E}_\mu$ — математическое ожидание по мере $\mathbf{P}_\mu$, порожденной наблюдениями (3), $\|\cdot\|$ и $\langle\cdot,\cdot\rangle$ обозначают норму и скалярное произведение в $\mathbb{R}^n$ соответственно, то есть $\|x\|^2 = \sum_{i=1}^n x_i^2, \quad \langle x, y\rangle = \sum_{i=1}^n x_i y_i$.

Нетрудно показать, что средне-квадратичный риск линейной оценки $\hat{\mu}^h(Y)$ вычисляется следующим образом:

$$R(\hat{\mu}^h, \mu) = \|(1 - h) \cdot \mu\|^2 + \sigma^2 \|h\|^2,$$

где $x \cdot y$ означает покоординатное произведение векторов $x, y \in \mathbb{R}^n$, то есть $x \cdot y \in \mathbb{R}^n$ с компонентами $x_i y_i$, $i = 1, \dots, n$.

Очевидно, что этот риск зависит от $h$ и мы можем найти его минимум по $h \in \mathcal{H}$, который часто называют оракульным риском

$$r^{\mathcal{H}}(\mu) = \min_{h \in \mathcal{H}} R(\hat{\mu}^h, \mu).$$

Однако, мы не можем использовать оценку

$$\mu^*(Y) = h^* \cdot Y, \quad h^* = \arg\min_{h \in \mathcal{H}} R(\hat{\mu}^h, \mu),$$

так как она зависит от неизвестного вектора $\mu$. Поэтому нашей целью является построение оценки $\tilde{\mu}^{\mathcal{H}}(Y)$ на основе заданных оценок $\hat{\mu}^h(Y),\ h \in \mathcal{H}$, такой, чтобы ее риск был как можно ближе к риску оракула. То есть мы хотим, чтобы равномерно по $\mu \in \mathbb{R}^n$ выполнялось неравенство следующего вида:

$$R(\tilde{\mu}^{\mathcal{H}}, \mu) \leq r^{\mathcal{H}}(\mu) + \tilde{\Delta}^{\mathcal{H}}(\mu),$$

где $\tilde{\Delta}^{\mathcal{H}}(\mu)$ — остаточный член, который мы хотели бы сделать меньше риска оракула $r^{\mathcal{H}}(\mu)$. Такого типа неравенства часто называются оракульными. Основной задачей является поиск оценок с минимальным остаточным членом. В общем случае эта задача не имеет решения. Однако в некоторых случаях удается построить оценку $\tilde{\mu}^{\mathcal{H}}(Y)$ такую, что:

- $\tilde{\Delta}^{\mathcal{H}}(\mu) \leq \tilde{C} r^{\mathcal{H}}(\mu)$ для всех $\mu \in \mathbb{R}^n$, где $\tilde{C} > 1$ константа,
- $\tilde{\Delta}^{\mathcal{H}}(\mu) \ll r^{\mathcal{H}}(\mu)$ для всех $\mu :\ r^{\mathcal{H}}(\mu) \gg \sigma^2$.

Известно, что можно построить оценку с приведенными выше свойствами, если векторы в $\mathcal{H}$ являются упорядоченными.

**Определение 1.** *Множество $\mathcal{H}$ состоит из упорядоченных векторов, если*

- $h_i \in [0,1],\ i = 1, \ldots, n$ *для всех* $h \in \mathcal{H}$,
- $h_{i+1} \leq h_i,\ i = 1, \ldots, n$ *для всех* $h \in \mathcal{H}$,
- *если для некоторого натурального $k$ и некоторых $h, g \in \mathcal{H}$*

$$h_k < g_k, \quad \textit{тогда} \quad h_i \leq g_i \quad \textit{для} \quad i = 1, \ldots, n.$$

Последнее условие означает, что векторы из $\mathcal{H}$ могут быть естественным образом упорядочены, так как для любых $h, g \in \mathcal{H}$ возможно только два случая: $h_i \leq g_i$ или $h_i \geq g_i$ для всех $i = 1, \ldots, n$.

В качестве мотивации для задачи оценивания зашумленного вектора при заданном множестве упорядоченных оценок рассмотрим задачу оценивания регрессионной функции с помощью сглаживающих сплайнов. Необходимо восстановить гладкую функцию $f(x),\ x \in [0,1]$ по наблюдениям

$$Z_i = f(X_i) + \epsilon\xi_i', \quad i = 1, \ldots, n, \tag{5}$$

где $X_i \in (0,1)$ и $\xi_i'$ независимые случайные величины со стандартным нормальным распределением. В качестве оценок функции регрессии $f(x),\ x \in [0,1]$ будем использовать сглаживающие сплайны, которые определяются как решения следующей оптимизационной задачи

$$\hat{S}_p(x, X, Z) = \arg\min_f \left\{ \frac{1}{n} \sum_{i=1}^{n} [Z_i - f(X_i)]^2 + p \int_0^1 [f^{(m)}(x)]^2 \, dx \right\}, \tag{6}$$

где $f^{(m)}(\cdot)$ — производная порядка $m$ и $p > 0$ — сглаживающий параметр.

Для того чтобы свести задачу оценивания функции регрессии с помощью сплайнов к задаче восстановления вектора в белом гауссовском шуме (3), перейдем в базис Райнша–Деммлера $\psi_k(x),\ x \in [0,1],\ k = 1, \ldots, n$. Этот базис обладает свойством двойной ортогональности

$$\langle \psi_k, \psi_l \rangle_n = \delta_{kl}, \quad \int_0^1 \psi_k^{(m)}(x) \psi_l^{(m)}(x) \, dx = \delta_{kl} \lambda_k,\ k, l = 1, \ldots, n, \tag{7}$$

где здесь и ниже $\langle u, v \rangle_n$ обозначает скалярное произведение

$$\langle u, v \rangle_n = \frac{1}{n} \sum_{i=1}^{n} u(X_i) v(X_i),$$

и собственные числа $\lambda_k$ упорядочены $\lambda_n \geq \ldots \geq \lambda_1$. Функцию $f(\cdot)$ и наблюдения $Z$ можно разложить по базису Райнша-Деммлера следующим образом:

$$f(X_i) = \sum_{k=1}^{n} \psi_k(X_i) \mu_k, \quad Y_k = \langle Z, \psi_k \rangle_n = \mu_k + \frac{\epsilon}{\sqrt{n}} \xi_k. \tag{8}$$

Затем, подставляя (8) в (6), приходим к

$$\hat{S}_p(x, X, Y) = \sum_{k=1}^{n} \hat{\mu}_k \psi_k(x),$$

где

$$\hat{\mu}_k = \arg\min_{\mu}\left\{\sum_{k=1}^{n}[Y_k - \mu_k]^2 + p\sum_{k=1}^{n}\lambda_k\mu_k^2\right\} = \frac{Y_k}{1+p\lambda_k}.$$

Поэтому задачи (3)-(4) и (5)-(6) эквивалентны при $\sigma = \epsilon/\sqrt{n}$, причем

$$\mathcal{H} = \left\{h : h_k = \frac{1}{1+p\lambda_k},\ p > 0\right\}.$$

**Во втором разделе** приводится доказательство оракульного неравенства для метода агрегации упорядоченных оценок с помощью экспоненциального взвешивания в задаче оценивания зашумленного вектора.

Рассматривается оценка, полученная с помощью экспоненциального взвешивания оценок $\hat{\mu}^h,\ h \in \mathcal{H}$, то есть оценка следующего вида

$$\bar{\mu}(Y) = \sum_{h\in\mathcal{H}} w^h(Y)\hat{\mu}^h(Y), \tag{9}$$

где $w^h(Y)$ положительные веса, такие что $\sum_{h\in\mathcal{H}} w^h(Y) = 1$ и

$$w^h(Y) = \pi^h \exp\left[-\frac{\bar{r}(Y,\hat{\mu}^h)}{2\beta\sigma^2}\right] \Big/ \sum_{g\in\mathcal{H}} \pi^g \exp\left[-\frac{\bar{r}(Y,\hat{\mu}^g)}{2\beta\sigma^2}\right], \quad \beta > 0,$$

$\bar{r}(Y,\hat{\mu}^h)$ — несмещенная оценка риска линейной оценки $\hat{\mu}^h(Y)$, а именно

$$\bar{r}(Y,\hat{\mu}^h) = \|(1-h)Y\|^2 + 2\sigma^2\sum_{i=1}^{n} h_i - \sigma^2 n.$$

Априорные веса $\pi^h$ зададим следующим образом

$$\pi^h = \left[1 - \exp\left\{-\frac{\|h^+\|_1 - \|h\|_1}{\beta}\right\}\right],$$

где

$$h^+ = \min\{g \in \mathcal{H} : g > h\},$$

$\pi^{h_{\max}} = 1$, где $h^{max} = \max_{h\in\mathcal{H}} h$, и $\|h\|_1 = \sum_{i=1}^{n}|h_i|$.

Также будем предполагать выполненным следующее условие.

**Условие 1.** *Существует постоянная $K_\circ \geq 0$, такая что для всех $h \geq g$ из $\mathcal{H}$*

$$\|h\|^2 - \|g\|^2 \geq K_\circ\big(\|h\|_1 - \|g\|_1\big).$$

Следующая теорема является основным результатом раздела.

**Теорема 5.** *Пусть $\beta \geq 4$ и выполнено Условие 1. Тогда равномерно по $\mu \in \mathbb{R}^n$ выполнено следующее неравенство:*

$$\mathbf{E}_\mu \|\bar{\mu} - \mu\|^2 \leq r^{\mathcal{H}}(\mu) + 2\beta\sigma^2 \log\left[C\left(1 + \frac{r^{\mathcal{H}}(\mu)}{\sigma^2}\right)\right],$$

*где $C = C(K_\circ, \beta)$ строго положительная постоянная, зависящая от $K_\circ, \beta$.*

Заметим, что этот результат улучшает классическое неравенство Кнайпа.

**В третьем разделе** рассматривается следующее обобщение задачи оценивания зашумленного вектора при помощи агрегации упорядоченных оценок из заданного множества. Задача состоит в том, чтобы по наблюдениям

$$Y_i = \mu_i + \sigma\xi_i, \quad i = 1, 2, \ldots, n, \tag{10}$$

где $\xi_i$ — стационарный гауссовский процесс с нулевым средним и $\mathbf{E}\xi_i^2 = 1$, оценить неизвестный вектор $\mu = \{\mu_1, \ldots, \mu_n\}$. Основная цель состоит в том, чтобы показать, что оракульные неравенства для метода экспоненциального взвешивания упорядоченных линейных оценок, полученные для белого шума, останутся справедливыми и для цветных шумов.

В качестве мотивации задачи (10) рассмотрим оценивание функции регрессии в гетероскедастичном шуме с помощью сглаживающих сплайнов. В этой задаче необходимо оценить функцию $f(x),\ x \in [0, 1]$ по зашумленным наблюдениям

$$Z_i = f(X_i) + \sigma(X_i)\zeta_i, \quad i = 1, \ldots, n, \tag{11}$$

где $\zeta$ — стандартный белый гауссовский шум, а $\sigma^2(x),\ x \in [0, 1]$ — неизвестная непрерывная функция. В качестве оценок будем использовать сглаживающие сплайны, которые являются решением оптимизационной задачи (6). Для того чтобы проверить эквивалентность моделей (11) и (10), воспользуемся базисом Райнша-Деммлера $\psi_k(x),\ x \in [0, 1],\ k = 1, \ldots, n$ из (7). Представляя функцию

регрессии в виде $f(x) = \sum_{k=1}^{n} \psi_k(x)\mu_k$, получаем, что наблюдения (11) эквивалентны

$$Y_k = \langle Z, \psi_k \rangle_n = \mu_k + \xi'_k, \quad \text{где} \quad \xi'_k = \sum_{j=1}^{n} \sigma(X_i)\psi_k(X_i)\zeta_i.$$

При равномерном распределении точек $X_i$ для базиса Райнша-Деммлера известна следующая асимптотика:

$$\psi_k(x) \approx \sqrt{\frac{2}{n}} \cos(\pi k x), \quad n, k \to \infty.$$

Поэтому при больших $n, k, j$

$$\mathbf{E}\xi'_j\xi'_k \approx \frac{1}{n}\sum_{p=1}^{n} \sigma^2(X_i)\cos[\pi(k-j)p].$$

Таким образом, гауссовская последовательность случайных величин $\xi'_k$, $k = 1, \ldots, n$ близка к стационарной. При этом для дисперсии шума в (10) справедлива асимптотическая формула $\sigma^2 \approx \sum_{i=1}^{n} \sigma^2(X_i)/n$. Отметим также, что эту величину несложно оценить по наблюдениям (11), например, с помощью оценки $\bar{\sigma}^2(Z) = \frac{1}{2n}\sum_{i=1}^{n-1}[Z_i - Z_{i+1}]^2$. Поэтому параметр $\sigma$ в (10) можно считать известным.

Оказывается, что для этой задачи оценивания вектора в стационарном гауссовском шуме с помощью метода экспоненциального взвешивания (9) верен аналогичный Теореме 5 результат.

**Теорема 6.** *Пусть $\beta \geq 4$ и выполнено Условие 1. Тогда равномерно по $\mu \in \mathbb{R}^n$ выполнено следующее неравенство:*

$$\mathbf{E}_\mu \|\bar{\mu} - \mu\|^2 \leq r^{\mathcal{H}}(\mu) + 2\beta\sigma^2 \log\left[C\left(1 + \frac{r^{\mathcal{H}}(\mu)}{\sigma^2}\right)\right],$$

*где $C = C(K_\circ, \beta)$ строго положительная постоянная, зависящая от $K_\circ, \beta$.*

**В третьей главе** приведены результаты вычислительных экспериментов. Первый раздел посвящен сравнению методов одномерной интерполяции данных. Во втором разделе сравнивается предложенный метод контроля точности для интерполяции сплайнами с крикингом. По результатам экспериментов для гладких

функций контроль точности интерполяции, который обеспечивают сплайны, оказался существенно лучше того, что дает кригинг. В третьем разделе представлены результаты сравнения рисков агрегированной оценки сглаживающих сплайнов для разных значений параметра $\beta$ в методе экспоненциального взвешивания. Получено, что в случае, когда отношение сигнал/шум невелико, экспоненциальное взвешивание позволяет получить оценку с существенно меньшим риском.

**В заключении** перечислены следующие основные результаты.

1. Доказана эквивалентность интерполяционных сплайнов и наилучшей интерполяции для обобщенных гауссовских процессов со специальными спектральными плотностями. Вычислена величина минимаксной ошибки интерполяции в классе гладких процессов.

2. Вычислена минимаксная ошибка интерполяции гладких функций из соболевских классов, заданных значениями на равномерной решетке со случайным равномерным сдвигом. Показано, что эта ошибка близка к минимаксной ошибке интерполяции на классе гладких процессов.

3. На основе эквивалентности сплайнов и наилучших интерполяций обобщенных гауссовских процессов предложен метод контроля точности интерполяции сплайнами. Показано, что для определенного класса процессов предложенный метод является хорошей оценкой для реальной ошибки.

4. Разработаны новые подходы к доказательству оракульных неравенств для метода экспоненциального взвешивания. С их помощью получены оракульные неравенства в задаче оценивания функции регрессии с помощью сглаживающих сплайнов.

5. Проведены численные эксперименты, которые показали, что в случае, когда отношение сигнал/шум невелико, экспоненциальное взвешивание позволяет получить оценку с существенно меньшим риском.

## Список публикаций


1. Голубев Г. К., Крымова Е. А. Об интерполяции гладких процессов и функций // Проблемы Передачи Информации. 2013. Т. 49. С. 61–84.

2. Крымова Е. А., Черноусова Е. О. Оракульное неравенство для метода экспоненциального взвешивания упорядоченных оценок // Труды МФТИ. 2013. Т. 5, № 3(19). С. 55–66.

3. Chernousova E., Golubev Y., Krymova E. Ordered smoothers with exponential weighting // Electronic Journal of Statistics. 2013. Vol. 7. Pp. 2395–2419.

4. Голубев Г. К., Крымова Е. А. Сплайны и стационарные гауссовские процессы // Доклады 9-ой Международной конференции «Интеллектуализация обработки информации». 2012. С. 207–211.

5. Голубев Г. К., Крымова Е. А. Splines and stationary Gaussian processes // Информационные технологии и системы – 2012 (ИТиС 2012): сб. трудов конференции. ИППИ РАН, 2012. С. 145–150.

6. Крымова Е. А. Оракульное неравенство для метода экспоненциального взвешивания упорядоченных оценок // Труды 55-й научной конференции МФТИ. МФТИ, 2012. С. 147–149.

7. Chernousova E., Golubev Y., Krymova E. On oracle inequality for exponential weighting of ordered smoothers // Proceedings of the 18th European Young Statisticians Meeting (EYSM- 2013). 2013. Pp. 1–5.

8. Krymova E. Oracle inequalities for the exponential weighting method in the case of regression estimation problem // Информационные технологии и системы – 2013 (ИТиС 2013): сб. трудов конференции. ИППИ РАН, 2013. С. 348–351.

Крымова Екатерина Александровна

Сплайны в задачах интерполяции и регрессионного анализа гауссовских процессов и гладких функций.

АВТОРЕФЕРАТ

Подписано в печать 12.11.2013.

Формат 60 × 84 1/16. Усл. печ. л. 1,0. Тираж 100 экз. Заказ 354.

Федеральное государственное автономное образовательное учреждение высшего профессионального образования «Московский физико-технический институт (государственный университет)»

Отдел оперативной полиграфии «Физтех-полиграф»

141700, Московская обл., г. Долгопрудный, Институтский пер., 9.